Deep Convolutional Neural Network for Voice Liveness Detection

Siddhant Gupta, Kuldeep Khoria, Ankur T. Patil, and Hemant A. Patil
Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India
E-mail: {siddhant_gupta, kuldeep_khoria, ankur_patil, hemant_patil}@daiict.ac.in

***Abstract*—In this work, we present the system to detect the liveness by identifying the pop noise in the voice signal in order to avoid the security breach of ASV systems. Pop noise is created due to spontaneous breathing while uttering a certain phonemes, and it has low-frequency characteristics. Given the low-frequency characteristics of the pop noise, we have used the short-time Fourier transform (STFT) with low-frequency contents (0-40 Hz) as a feature set along with a convolutional neural network as a classifier. The experiments are performed using the recently released *POp noise COrpus* (POCO) dataset. We have considered the approach given in the original POCO dataset paper as a baseline and compared the results with the proposed architecture. The performance of the proposed architecture is measured using *10*-fold cross-validation and the customized disjoint partition of the dataset. It is observed that the proposed architecture shows an improvement in accuracy for voice liveness detection in both cases. In particular, the proposed architecture obtained 19.86% and 18.22% absolute improvement in accuracy for *10*-fold cross-validation and customized data partition, respectively, as compared to the baseline.**

Index Terms: Voice liveness detection, Pop noise, CNN, POCO dataset.

I. Introduction

In recent decades, biometric authentication has gained significant traction in many areas, such as financial transactions, personalized devices, such as mobile phones and laptops, electronic applications. There are various choices of the biometrics, such as fingerprint, palm, iris, gait, face, and voice [1]. Among these, voice biometric is more natural way of communication with the machines and hence, it has emerged in many real-time applications, such as voice assistants, controlling Internet of Things (IoT) devices, etc. This became possible due to development of robust automatic speaker verification (ASV) technology [2]. However, due to parallel advancement in the other speech technologies, such as speech synthesis (SS) and voice conversion (VC), ASV systems have become susceptible to spoofing attacks [3], [4], [5], [6], [7]. Furthermore, high quality recording and playback devices facilitate the replay spoofing attack, which became difficult case for ASV to identify spoofing [8], [9]. To alleviate the issue of spoofing attacks on the ASV system, ASVspoof challenge campaigns were initiated in 2015, 2017, and 2019, which were organized as special sessions during INTERSPEECH conferences [10], [11], [12]. These campaigns distributed the datasets, protocols, and evaluation metrics to be able to provide the common platform for comparison of the performances of various countermeasure (CM) systems. These campaigns encompasses the SS, VC, and replay spoofing attacks. In this study, we propose to develop the CM system for Voice Liveness Detection (VLD) in the context of ASV. The relationship between the presence of pop noise in recorded speech and distance between speaker and microphone is inversely proportional. Moreover, the microphone can sense the breathing noise, i.e., pop noise of the speaker if the distance is very small (approx. 5 cm). Hence, pop noise can be attributed to the live genuine speech.

To the best of authors' knowledge, liveness detection for anti-spoofing is proposed for the first time in [13], where two approaches of liveness detection are proposed: (a) low-frequency-based single channel detection, (b) subtraction-based pop noise detection with two channels. In the former approach, Short-Time Fourier Transform (STFT) around lower frequency region is utilized (as the pop noise exists in the lower frequency regions). Whereas in the later approach, entire frequency range of the spectrum is utilized. In [14], phoneme-based pop noise detection is performed for VLD along with ASV system, where pop noise duration is detected in an utterance and estimated phonemes in that duration are analyzed for VLD task. The similar approach of phoneme-based pop noise detection was utilized in [15] with extended study on Gammatone Frequency Cepstral Coefficients (GFCC) feature set for pop noise detection.

During natural speech production, airflow travels from the lungs to the vocal folds, excites vocal tract system and finally, bursts out from the mouth as a sound wave. While capturing this sound via microphone, if the distance between speaker and microphone is small, the microphone in addition to capturing speech signal, can also capture the friction between the lips as *bursts* which is termed as *pop noise*. The intensity of this *pop noise* detected by the microphone is inversely proportional to the distance between the speaker and the microphone. Such *pop noise* will not be recorded if the recorder is kept far away from the speaker. An attacker who is deceptively trying to record the voice usually may not be able to put the recording device near to the speaker which will result in the absence of *pop noise* from the recorded speech. Hence, *pop noise detection* can provide reliable acoustic cues for VLD and thus, should be able to distinguish between the live (genuine) speech and the replayed speech [16].

Recently, POCO dataset is developed which can be used to build the system for VLD by identifying the pop noise which causes the distortion in the speech signal introduced by the speaker's breath [17]. Thus, pop noise is the characteristic

of the live speech. Identifying the pop noise for live speaker detection might be very useful strategy in the applications, where the testing microphone is placed at a short distance from the speaker, and consequently this strategy may protect the ASV system from the spoofing attacks. The architecture proposed in [13] is the popular approach for VLD and consequently, it is used to produce the results in original POCO dataset study in [17]. With this reference study, we considered this approach as a baseline approach. Our proposed system uses low-frequency regions in Short-Time Fourier Transform (STFT) spectrogram as a feature set along with convolutional neural network (CNN) as a classifier.

The rest of the paper is organized as follows: Section II presents the details of the proposed approach whereas Section III gives POCO dataset details. Section IV provides details of the experiments and the results obtained. Finally, Section V summarizes our work along with some limitations of current methods and future research directions.

II. Proposed Approach

A. Baseline Algorithm

In our work, detection of pop noise is considered as binary classification task, where utterances with and without pop noise are labelled as 1 and 0, respectively. Spectrograms are used as input features. The baseline is implemented using the methodology from [17]. Let $x(n)$ be the input signal. STFT is calculated as :

$$\begin{aligned} X(\omega,\tau) &= \sum_{n=-\infty}^{\infty} x(n)\cdot w(n,\tau)\cdot e^{-j\omega n}, \\ &= \sum_{n=-\infty}^{\infty} x(n,\tau)\cdot e^{-j\omega n}, \end{aligned} \tag{1}$$

where $x(n,\tau) = x(n)\cdot w(n,\tau)$ is the windowed speech segment centered at τ. Now, spectrogram (spectral energy density) is obtained by calculating the magnitude square of $X(\omega,\tau)$, i.e.,

$$S(\omega,\tau) = |X(\omega,\tau)|^2, \tag{2}$$

$S_{eng}(\omega,\tau)$ is calculated by considering spectral energy density from $S(\omega,\tau)$ ranging within the frequency bins corresponding to [0, ω_{max}], i.e.,

$$S_{eng}(\omega,\tau) = |S(\omega,\tau)|_{0\le\omega\le\omega_{max}}. \tag{3}$$

Since the pop noise is observed in the lower frequency region of the spectrogram features, ω_{max} is the frequency in rad/sec corresponding to $40Hz$. ω_{avg} is calculated as the average of the spectral energies for each frame. Then, the mean and standard deviation is taken for ω_{avg} across all the bins. This results in $1\times N$ vector, where N is number of frames. Next, $\omega_{avg(i)}$ is considered for each of the i^{th} frame which is calculated as :

$$\omega_{\text{avg(i)}} = \frac{1}{N_b}\sum_{\omega=0}^{40\ Hz} |S_{eng,i}(\omega,\tau)|. \tag{4}$$

where N_b represents number of frequency bins upto 40 *Hz*. Then, mean and standard deviation is estimated for averaged spectral energies $\omega_{avg(i)}$ in order to normalize it. Then, 10 frame indices with largest spectral energies were taken from the normalized $\omega_{avg(i)}$, and frames corresponding to those indices were chosen from $S_{eng}(\omega,\tau)$. This feature set with appropriate labels is fed to Support Vector Machine (SVM) for classification purpose. Further details of this baseline algorithm can be found in [17].

B. Proposed Algorithm

We propose a deep learning-based approach for the detection of pop noise. In our work, CNN is used for the classification of pop noise. We have used spectral energy densities of the spectrogram as input features to the CNN classifier. The reason behind opting CNN as a classifier is that it captures the presence of pop noise in the spectrogram more predominantly when compared to the SVM classifier as there is a significant change in spectral energy at the pop noise instances in the spectrogram. Moreover, to ensure that the learning of the model is done on the basis of pop noise effect, we have considered spectral energy densities of the spectrogram S_{eng} only for the lower frequency regions, i.e., $0-40Hz$. The window length for obtaining the spectrogram was set as 25 ms and a hop size of 4 ms. We have considered a frequency resolution of 1 Hz which result in 40 bins corresponding to 40 Hz. Furthermore, S_{eng} is modified to give a matrix size of 40x400 by clipping the feature maps whose size was more than 40x400 or by padding the feature maps by concatenating the data from the same feature map to get the required size. This provides us a uniform size matrix to feed the CNN classifier having 40 frequency bins and 400 frames. Z-normalization is used to normalize the utterances. These modified feature maps, all of sizes 40x400, are used as the input to our neural network.

The CNN network consists of 3 convolution blocks (referred to as Convolution 1, Convolution 2, and Convolution 3 in Fig. 1), and 3 Fully-Connected (FC) layers (referred to as FC 1, FC 2, and FC 3 in Fig. 1). Each convolution block consists of a $2-D$ convolution layer followed by a max-pooling layer to remove the inconsistencies in the feature map. Both convolution and max-pooling operations are done using kernel size of 3x3. In addition, convolution operation is performed using zero padding with a stride of 1. The final convolution block is followed by 3 fully-connected linear layers with different hidden units. The output of the final layer is activated using a sigmoid function, which makes the final decision of whether the utterance contains pop noise or not. Rectified Linear Unit (ReLU) function is used as the activation function in the hidden layers.

The model is trained using Stochastic Gradient Descent (SGD) algorithm with a batch size of 64, and learning rate of 0.001. Binary cross-entropy loss is chosen as the loss function. The experiments are executed for a total number of 400 epochs. The experiments are performed using speaker-independent *10-fold* cross-validation strategy and customized

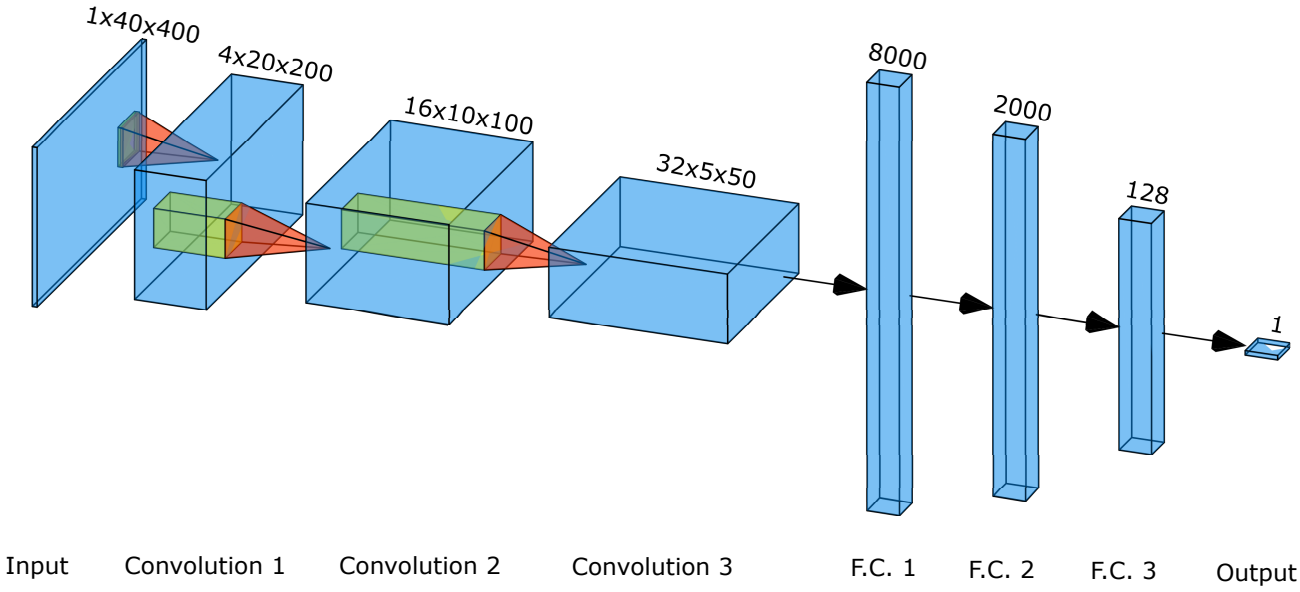

Fig. 1. The customized Convolutional Neural Network (CNN) architecture for the liveness detection task. After [18].

disjoint partition of the dataset as shown in Table I.

III. Details of POCO Dataset

In realistic scenarios, if an attacker tries to attempt a spoofing attack, he/she must somehow obtain the voice samples of the target (genuine) speaker. The simplest way to do this is by recording (eavesdropping) the voice of target speaker and then replaying it infront of the ASV system. Since these recordings will be done from long distances, pop noise will not be recorded by the attacker's microphone and this absence of pop noise in the replayed speech will be able to flag the spoofed speech from the genuine speech.

In this work, we have used recently released *POCO* dataset [17]. There are a total of 66 speakers out of which 34 are male and 32 are female. The words were selected from the English language such that all the 44 phonemes are covered in the recording. The dataset is sampled at 22050 Hz sampling frequency with a bit-depth of 24-bits. The dataset has three subsets, namely, RC-A (Recording with Microphone), RP-A (Eavesdropping), and RC-B (Recording with Microphone Array). We have excluded the RC-B subset for our experiments as it consists of microphone array, and it's corresponding spoof speech utterances are not provided. In addition, the experiments in [17] are performed using RC-A and RP-A subsets. The details of RC-A and RC-B are as follows:

A. Recording with Microphone (RC-A)

This subset represents genuine speaker as it was recorded directly with the live speaker and hence, contains pop noise. The recording was done with Audio-Technica AT4040 microphone. The distance between speaker and microphone was fixed to be 10 cm.

B. Eavesdropping (RP-A)

Eavesdropping is done to imitate a scenario where replay attack is done by an attacker from a long distance, i.e., without pop noise. This scenario is simulated by using Audio-Technica AT4040 microphone with a pop filter inserted between speaker and microphone. The distance between speaker and microphone was fixed as 10 cm.

TABLE I
Statistics of the POCO dataset for our experiments

Subset	# Utterances	# Speaker	# Male	# Female
Training	13552	53	26	27
Evaluation	3432	13	6	7

The dataset is partitioned into training and evaluation subsets as 80% and 20% utterances, respectively. Each of these subsets consist of half of the genuine and half of the spoof speech utterances. We also ensured that the speakers are exclusive in each subset and the ratio between male and female speaker is maintained. The statistics of the data distribution in training and evaluation subset is shown in Table I.

IV. Experimental Results

A. Spectrographic Analysis

Panel I of Fig. 3 represents the speech signal and it's corresponding spectrograms for the word, 'thong' for genuine speech, whereas Panel II shows similar plots for spoofed speech. There is a presence of high spectral energy density at low frequency region for genuine speech (Panel I(b)), which is not observed in the spectrogram of spoofed speech (Panel II(b)). Thus, the pop noise is present for genuine speech at low frequency regions and is absent for spoofed speech. This spectrographic difference is very well captured by the CNN to do the classification of genuine *vs.* spoofed speech.

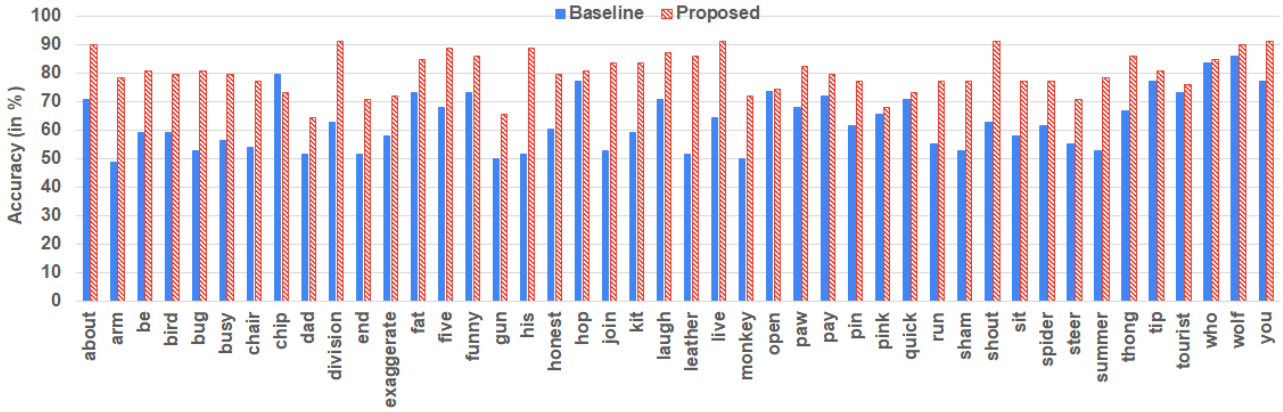

Fig. 2. Comparison of word accuracy on test data for baseline *vs.* proposed algorithm.

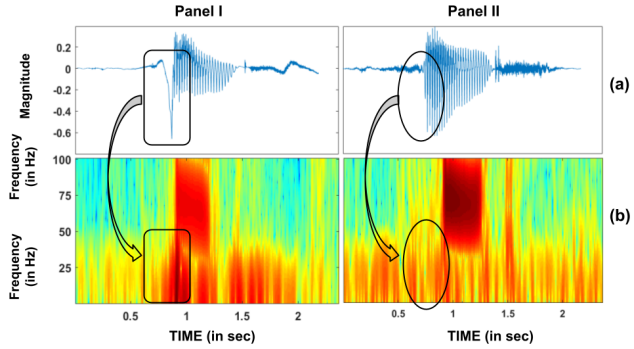

Fig. 3. Spectrograms for word 'thong'. Panel I.(a) and (b) show the speech signal and spectrogram for the genuine utterance. Panel II.(a) and (b) show similar plots for spoofed utterance. The pop and non-pop locations are highlighted by rectangle and circle box, respectively, for both the spectrograms.

B. *Word-wise Performance*

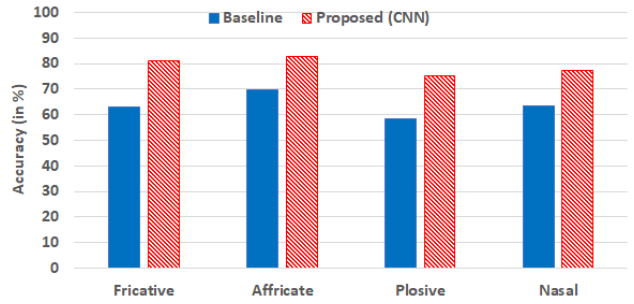

Fig. 4. Average accuracy (in %) for different types of sounds.

We have obtained an overall accuracy of 80.51% when the training and evaluation was done according to the data mentioned in Table I for proposed CNN-based approach, as compared to overall accuracy of 62.29% for the baseline algorithm. In addition, we have also performed *10*-fold cross-validation and obtained overall accuracy of 82.15% for the proposed approach. It can be observed that the absolute improvement of 18.22% and 19.86% is obtained by the proposed algorithm over the baseline for evaluation done using customized dataset and *10*-fold cross-validation, respectively. The experiments are also performed by increasing the ω_{max} in eq. (3) upto 100 Hz. However, we obtained relatively better performance for ω_{max} = 40 Hz.

In addition, accuracy is also analyzed for the individual words in the dataset. Fig. 2 represents the word-wise accuracy for baseline *vs.* proposed algorithm on the evaluation set. Here, we can observe that for words which have the higher probability of presence of pop noise (such as, 'division', 'fat', 'funny', 'five', 'thong', 'shout', 'wolf', 'you'), our proposed architecture performs relatively better than the baseline. The proposed architecture shows the average accuracy of 85 % for these words as opposed to 70 % accuracy for the baseline. This 15 % absolute improvement in average accuracy for these words indicate the capability of our proposed architecture for pop noise detection. For the other words, our approach still outperforms the baseline, though with comparatively lower improvement in accuracy. We have also analyzed the performance of both the approaches on different types of speech sounds. Fig. 4 shows the plot for average accuracy for fricative, affricate, plosive, and nasal sounds. Here, it can be observed that average accuracy for fricative and affricate sounds is higher (around 65 %) for the baseline and 80 % for proposed approach than that for plosive and nasal sounds (i.e., around 60 % for baseline and 75 % for the proposed approach). This is an expected result because fricative and affricate will have higher accuracy as there is higher probability of presence of pop noise in these types of speech sounds [19]. This is because during the production of fricative, the air turbulence creates a chaotic mix of random frequencies which lasts for very short time due to which there is momentary burst of energy occurring at random frequencies, and since pop noise is perceived as the air burst, pop noise characteristics are obtained predominantly for fricative sound. In addition, affricate is produced by stopping the airflow initially in the vocal tract system and then releasing it in the same manner as in fricative and hence, presence of pop noise is also captured for affricate sound. On the other hand,

while producing plosive sounds, there is a release of burst which looks like a very thin fricative and hence, as the burst duration is very thin when compared to the fricative, detecting pop noise in plosive sounds is difficult and hence, average accuracy is relatively lower for plosive sounds. Whereas, in case of nasal sounds, spectral energy density is mostly cohenrent at lower frequency region [19] and also suffer from poor spectral resolution caused by higher $-3dB$ bandwidth of nasal cavity, as its impulse response gets quickly damped due to its highly complex surface of the nasal cavity [19]. Hence, it may be possible that both the algorithms are *confused* as to whether the speech contains actual pop noise or the nasal sound resulting in significant misclassification.

V. Summary and Conclusions

In this paper, we have used pop noise as an indicator of the liveness of genuine speech and its ability to be used as a potential acoustic cue for replay spoof attack detection in the context of ASV. For VLD task, we proposed a novel deep learning-based approach for the classification of genuine *vs.* spoofed speech. Spectrogram features extracted from the recently published POCO dataset were used as input to the baseline and our proposed approach. It was observed that our proposed approach provides significantly better results indicating that CNN was able to learn the pop noise from the utterances successfully. However, our approach is based on the assumption that the distance between the speaker and the recorder is large enough so that no significant amount of pop noise can be captured. It can help in improving the robustness and sophistication of current voice privacy and spoof detection systems. In the future, it will be interesting to analyze the performance of more sophisticated neural network architectures for the pop noise detection task, which are explored in the ASVSpoof PA (physical access) dataset, and more baseline systems can be added for further research. Furthermore, the performance can be analyzed on different spectral feature sets for possible improvement in results.

Acknowledgment

The authors would like to thank the authorities at DA-IICT Gandhinagar for kind cooperation support during this research work.

References

[1] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," *Pattern Recognition Letters*, vol. 79, pp. 80–105, 2016.

[2] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "SdSV Challenge 2020: Large-Scale Evaluation of Short-Duration Speaker Verification," in *INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 731–735.

[3] Q. Tian, Z. Zhang, H. Lu, L.-H. Chen, and S. Liu, "FeatherWave: An Efficient High-Fidelity Neural Vocoder with Multi-Band Linear Prediction," in *INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 195–199.

[4] Y. Ai and Z.-H. Ling, "Knowledge-and-Data-Driven Amplitude Spectrum Prediction for Hierarchical Neural Vocoders," in *INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 190–194.

[5] P. chun Hsu and H. yi Lee, "WG-WaveNet: Real-Time High-Fidelity Speech Synthesis Without GPU," in *INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 210–214.

[6] S. Ding, G. Zhao, and R. Gutierrez-Osuna, "Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition," in *INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 776–780.

[7] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Recognition-Synthesis Based Non-Parallel Voice Conversion with Adversarial Learning," in *INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 771–775.

[8] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco *et al.*, "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, USA*, 2017, pp. 5395–5399.

[9] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVSpoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *INTESPEECH, Stockholm, Sweden*, 2017, pp. 2–6.

[10] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH, Dresden, Germany*, 2015, pp. 2037–2041.

[11] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: Meta-data analysis and baseline enhancements," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, 26 - 29 June, 2018.

[12] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *INTERSPEECH, Graz, Austria, 2019*, pp. 1008–1012.

[13] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in *INTERSPEECH, Dresden, Germany*, 2015, pp. 239–243.

[14] S. Mochizuki, S. Shiota, and H. Kiya, "Voice liveness detection using phoneme-based pop-noise detector for speaker verifcation," in *Odyssey 2018 The Speaker and Language Recognition Workshop. ISCA, Les Sables d'Olonne*, 2018, pp. 233–239.

[15] Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. Luo, "Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications, Paris, France*, 2019, pp. 2062–2070.

[16] S. Mochizuki, S. Shiota, and H. Kiya, "Voice livness detection based on pop-noise detector with phoneme information for speaker verification," *The Journal of the Acoustical Society of America (JASA)*, vol. 140, no. 4, pp. 3060–3060, 2016.

[17] K. Akimoto, S. P. Liew, S. Mishima, R. Mizushima, and K. A. Lee, "POCO: A voice spoofing and liveness detection corpus based on pop noise," *in INTERSPEECH, Shanghai, China*, pp. 1081–1085, 2020.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[19] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. 2^{nd} Edition, Pearson Education India, 2006.